Feature Affinity based Pseudo Labeling for Semi-supervised Person Re-identification

Guodong Ding, Shanshan Zhang, Salman Khan, *Member, IEEE*, Zhenmin Tang, Jian Zhang, *Senior Member, IEEE* and Fatih Porikli, *Fellow, IEEE*

***Abstract*—Vision-based person re-identification aims to match a person's identity across multiple images, which is a fundamental task in multimedia content analysis and retrieval. Deep neural networks have recently manifested great potential in this task. However, a major bottleneck of existing supervised deep networks is their reliance on large amount of annotated training data. Manual labeling for person identities in large-scale surveillance camera systems is quite challenging and incurs significant costs. Some recent studies adopt generative model outputs as training data augmentation.**

To more effectively use this synthetic data for an improved feature learning and re-identification performance, this work proposes a novel Feature Affinity-based Pseudo Labeling (FAPL) method with two possible label encodings. To the best of our knowledge, this is the first study that employs pseudo-labeling by measuring the affinity of unlabeled samples with the underlying clusters of labeled data samples using the intermediate feature representations from deep networks. We propose to train the network with the joint supervision of cross-entropy loss together with a center regularization term, which not only ensures discriminative feature representation learning but also simultaneously predicts pseudo-labels for unlabeled data. We show that both label encodings can be learned in a unified manner and help improve the overall performance. Our extensive experiments on three person re-identification datasets, Market-1501, DukeMTMC-reID and CUHK03, demonstrate significant performance boost over the state-of-the-art person re-identification approaches.

***Index Terms*—pseudo-labeling, semi-supervised learning, person re-identification, deep networks, generative modeling.**

I. Introduction

Person re-identification is one of the basic yet challenging visual understanding tasks in multimedia content analysis. It has a wide range of applications including search and retrieval [1], cross-camera tracking [2–4] and video summarization [5] to name a few. Given a query image, it searches a gallery set of images for detecting the instances of the same person depicted in the query image. The gallery often contains images acquired from different cameras, viewpoints, and modalities at different time instances [6, 7]. This causes large intra-class and small inter-class differences due to significant variations in illumination, object appearance, body posture, pose, imaging noise, blur, and occlusions [8].

In recent years, deep Convolutional Neural Networks (CNNs) have achieved remarkable success in person re-identification [8–11]. However, CNN based methods are limited by the insufficient amount of data available for each identity. Manual labeling is the main bottleneck for acquiring large-scale annotations due to the tedious and labor-intensive nature of the job. This problem is particularly more prominent in the case of person re-identification as the labeling involves manually selecting identities and associating images from different cameras with varying viewpoints, illumination, occlusions and body pose changes. This is evidenced by the fact that even recent large-scale datasets, such as Market-1501 and DukeMTMC-reID which have been acquired specifically for deep learning, still have very limited pedestrian images per identity. For example, Market-1501 dataset has on average 17.2 training images for 751 identities. Furthermore, the number of images is unevenly distributed such that some identities have as few as 2 samples, while only a few classes have more than 20 images. Therefore, it is highly important to perform intelligent data augmentation to extend the training set.

The emergence of Generative Adversarial Network (GAN) [12] has partially addressed this problem as they can generate novel images with good perceptual quality. However, a pressing issue is how to optimally use the synthetic data. Pseudo-labeling is then proposed as a technique to produce approximate labels for unlabeled data on the basis of labeled data instead of manually labeling them. Initial efforts towards this problem adopted simplistic approaches e.g., [13] created a single new label for all generated images while [14] used predictions from a pre-trained CNN model to label generated images. More recently, [15, 16] proposed to use Label Smooth Regularization (LSR) to assign pseudo-labels to synthetic data samples. LSR was proposed decades ago and revisited recently in [17] to reduce over-fitting by assigning small values (instead of 0) to non-ground-truth classes for cross-entropy loss computation. Specifically, [16] extends LSR to outliers (LSRO) by assigning uniformly distributed virtual labels to generated images from GAN networks. This choice was made to avoid emphatically classifying generated samples into one of the existing categories. Afterwards, [15] argued that generated images have considerable visual differences and assigning same labels to all would lead to ambiguous predictions. Thus, they proposed to provide labels based on the rankings of normalized class predictions (probability estimates) over all pre-defined classes.

Guodong Ding, Shanshan Zhang and Zhenmin Tang are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu, China. e-mails: guodong.ding@njust.edu.cn, shanshan.zhang@njust.edu.cn and tzm.cs@njust.edu.cn

Salman Khan and Fatih Porikli are with College of Engineering and Computer Science, Australian National University, Australia. e-mails: salman.khan@anu.edu.au, fatih.porikli@anu.edu.au

Jian Zhang is with School of Electrical and Data Engineering, University of Technology Sydney, Australia. e-mail: Jian.Zhang@uts.edu.au

One remarkable drawback of all existing pseudo labeling approaches is that they are agnostic to underlying relationships between an unlabeled and labeled data samples. The most mature effort in this direction [15] performs label generation based on class predictions that are solely dependent on the input sample and neglect distance based affinity between the unlabeled image and labeled examples. In this work, we attempt to overcome this shortcoming by dynamically associating unlabeled samples with pre-defined classes during the training process. Inspired by the spirit of clustering that leverages the underlying patterns within training data, we propose a novel label assigning approach called Feature Affinity based Pseudo Labeling (FAPL) which delivers significant performance boost in person re-identification. FAPL aggregates labeled data samples that belong to the same class into refined clusters and simultaneously provides pseudo-labels to unlabeled data samples based on their similarity with each cluster center in the feature space. Building upon feature affinities, we propose two possible labeling schemes i.e., one-hot pseudo-label based on non-maximum suppression and distributed pseudo-label for soft label assignment. The former is easier to implement and train, while the later one performs better in our evaluations.

Another observation that shed light on our work is that despite the one-hot and distributed label encodings derived from probability predictions have been proposed separately for unlabeled data, they fail to be combined together in a unified architecture. Previous efforts only considered unitary encoding, either one-hot or distributed. Specifically, [13, 14, 18] chose one-hot labels for unlabeled data, while [15, 16] adopted distributed labels. The reason behind this situation is that the guarantee of a valid network training procedure is effective weight gradients, however, using probability predictions directly as distributed labels can not provide any weight corrections. [15] is a workaround proposing to assign labels based on probability rankings which is somehow effective but inescapably introduces errors. In this work, we endeavor to address this problem by introducing feature affinities. Feature affinity itself is not reverent to any class probability, granting its ability to produce pseudo-label with both encodings in a unified way while ensuring effective learning process.

Considering the recent progress in adversarial networks, we also study the effect of using better generative models for pseudo-labeling. Several efforts have recently been devoted to enhance the visual quality of synthetic images and stabilize the model training process [12, 14, 19–23]. For this purpose, better loss functions have been explored as well as novel network configurations to generate realistic images. The use of extra information was investigated in Conditional GAN [19], where both generator and discriminator are conditioned on extra information, such as class labels, to improve the visual quality of generated samples. In this work, alongside DCGAN, we further experiment with the recent improved Wasserstein GAN (IWGAN) model which avoids modal collapse and generates high-quality samples with better convergence properties. The WGAN [22] is based on the Wasserstein distance measure as adversarial training loss function which is better suited to depict distances between distributions. Our experiments show that training with higher-quality images helps improve the re-identification performance.

The main contributions of this work are summarized as follows:

- A multi-task loss formulation is proposed for semi-supervised learning which has two advantages. First, it jointly considers inter-class and intra-class variations in feature space for more discriminative representation learning. Second, it can simultaneously estimate pseudo-labels for unlabeled data.
- We first propose to consider feature affinities between GAN generated samples and labeled data rather than prediction probabilities to estimate pseudo-labels with two possible encodings. Besides, both encodings can be generated uniformly based on feature affinities.
- Our experiments on three standard person re-identification benchmarks demonstrate that the proposed method achieves significant improvements over other pseudo-labeling approaches and also outperforms the best performing methods in most comparison cases.

The remainder of this paper is organized as follows. A review of related works is provided in Section II. Section III presents background knowledge on pseudo-labeling technique. In Section IV, we provide details of our proposed feature affinity based pseudo-labeling approach followed by a discussion on why FAPL works better. Section V exhibits the effectiveness of proposed methods on three standard person re-identification benchmarks and provides an extensive ablation study. We conclude our work in Section VI.

II. Related work

In this section, we discuss the relevant works on semi-supervised learning and person re-identification with deep network architectures, respectively.

A. Semi-supervised Learning

Semi-supervised learning uses both labeled and unlabeled data to improve performance on a given task. It is driven by its practical value in learning faster, cheaper, and better feature representations. In many real world applications, it is relatively easier to acquire a large amount of unlabeled data. Semi-supervised learning seeks to train a model that can make more accurate predictions on future unseen test data compared to a model learned only from labeled training data. Plenty of approaches have been proposed in the literature for this setting. Common semi-supervised learning methods include variants of generative models [24], graph Laplacian based methods [25], co-training [26], and multi-view learning [27]. Above works in semi-supervised learning are based on the fact that sufficient unlabeled data is available. However, collecting unlabeled data is also cumbersome in some applications. After the emergence of Generative Adversarial Network (GAN) [12], a branch of research on semi-supervised leaning has shifted to exploring GAN generated images [13, 14]. This work, incorporates unlabeled samples generated by GAN alongside the real samples available in the labeled datasets.

B. Person re-identification

Person re-identification aims at matching captured person images from different camera installations that belong to the same identity. Main efforts in this area can be divided into two categories: (a) metric learning and (b) feature representation learning. Metric learning usually takes input in the form of image pairs or triplets and learns a similarity metric using pairwise or triplet loss [28–30]. [30] proposes to use deep networks to learn a similarity metric directly from image pixel, sparing the trouble of feature crafting and engineering. [28] divides feature representations into four parts which are later concatenated for final triplet loss calculation. Metric learning has shown its great effectiveness in person re-identification, however, this stream of work suffers from huge data expansion when constituting image pairs and triplets especially when applied to large-scale datasets. [31] uses a part-based CNN to extract discriminative and stable feature representations and proposes a novel set-to-set (S2S) loss for similarity learning which ensures large margin between inter-class and intra-class set.

The other type of works focuses on feature learning, addressing this task in the form of classification. Common practices include first training a pedestrian identity predicting model and then extract last fully connected layer activations as pedestrian descriptor for retrieval during testing [8, 29, 32–34]. Amongst these, [33] proposes to feed global information into previous layer for a compact feature representation, [16] demonstrates that the use of DCGAN generated samples to enlarge the training set helps achieve a boost in performance. They propose a label smoothing regularization for unlabeled samples (called LSRO) to assign a distributed pseudo-label. A major drawback of their approach is the underlying assumption that the synthetic data does not belong to any class, therefore considering a uniform distribution for all unlabeled samples. Our work aims to address this limitation and proposes a novel loss function that automatically discovers patterns in the unlabeled data.

III. Background on Pseudo-Labeling

Manually labeling generated person images is impractical as the image quality can not be assured which would take huge effort to judge which identity it belongs to due to the variations in appearances. Thus, pseudo-labeling for generated pedestrian images becomes central. As mentioned above, there have been work adopting GANs to generate samples as a data augmentation solution in the field of person re-identification and investigating to find the optimal labeling scheme.

Existing pseudo-labeling approaches are depicted in Fig. 1 and summarized as follows:

- **All-in-one** [13, 14]. As illustrated in Fig. 1(a), all-in-one seeks the easiest solution to assign labels. It simply introduces an extra new class and directly groups all unlabeled data into it without considering any variations which may exist between all generated images. Unlabeled data are trained with this fixed new class label throughout the training procedure.

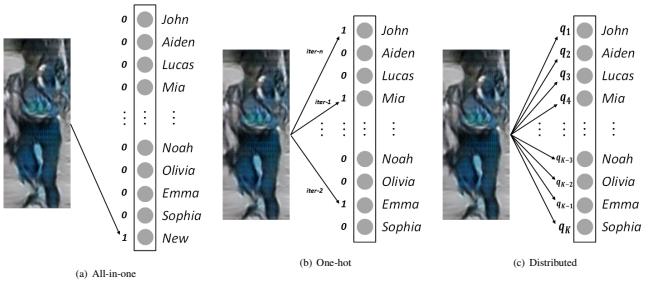

Fig. 1: Existing pseudo-labeling approaches can be divided into above three categories. From left to right are: All-in-one, One-hot and Distributed. All-in-one adds a new class label for all unlabeled data. One-hot assigns each unlabeled data a dynamic class label in each training epoch. Distributed considers contributions while labeling unlabeled data.

- **One-hot** [18]. On the basis of all-in-one, One-hot takes into consideration the intra-variations within all generated image and proposes to assume that each sample belongs to an existing class. Pseudo-labels are assigned by taking the maximum value for the probability prediction for each class shown as Fig. 1(b). One-hot label is identical in its form to ground-truth label, thus this pseudo-labeling scheme is easy to train. Note that as the training proceeds, pseudo-labels for the same image can be different as the predictions might change.
- **Distributed** [16], [15]. This type of pseudo-labeling further extends one-hot labeling scheme, and considers that the label for an unlabeled data should be distributed like displayed in Fig. 1(c). Since GAN generated images are fake samples drawn from the real data manifold, it would be inaccurate to classify them into any single class. Based on this assumption, [16] proposed to give equally distributed labels i.e., $q_i = 1/K, \quad i = 1, 2, ..., K$ in LSRO. In contrast, MpRL [15] assigns distributed labels according to class prediction ranks considering the class contributions. Distributed pseudo-labels together with real labels are trained with cross-entropy loss.

One proven ability of GAN is that it can generate samples from the training data distribution without strictly modeling it. In other words, GAN generated images can be seen as samples drawn from the labeled set. Thus, pseudo-labeling for generated images on the basis of labeled samples is more natural. However, none of existing pseudo-labeling methods takes into account the innate relations between labeled and unlabeled data to improve the feature representation learning under the semi-supervised framework. In contrast, this work aims to address this limitation and propose a novel loss function that automatically discovers patterns in the unlabeled data by associating them to labeled data samples. Next, we describe our semi-supervised learning approach based on pseudo-labeling.

IV. The proposed Approach

A. Overview

Despite the fact that generated data samples from GAN are unlabeled, they can be used alongside the labeled examples to improve the learned features representations in a semi-supervised setting. We propose to take into account the underlying patterns in the labeled data and leverage those to infer pseudo-labels for unlabeled data. To this end, we introduce a new multi-task learning objective for the semi-supervised training together with two new labeling schemes. The multi-task objective comprises of a normal classification loss and a center regularization term. The classification loss seeks to learn a ID-discriminative Embedding (IDE) for each pedestrian [35]. The center regularization term improves the discriminative ability of feature embedding and simultaneously predict pseudo-labels for generated samples.

The overall network architecture is illustrated in Fig. 2. The upper row denotes the training procedure of semi-supervised learning with synthetic images generated through GAN. It consists of two main modules. The **first** module on the left is the image generation module, where a generative model is optimized using adversarial training to estimate the data distribution based on existing real samples (e.g., labeled images from training set). The generator is trained to create samples that pass the discriminator test, whilst the discriminator is trained to separate fake samples from real ones. Afterwards, the trained generator is used to obtain large amounts of synthetic image samples lying on the approximated data manifold for subsequent training. The **second** module takes as input the generated data samples which are unlabeled. This module uses the unlabeled data samples alongside the labeled ones to learn feature representations with the joint supervision of classification and a center regularization term. The bottom row shows the testing phase, where activations output from the convolutional neural network (CNN) are used as pedestrian descriptors for a Euclidean distance based retrieval operation.

The objective function (illustrated in Fig. 2) of our proposed approach can be expressed as:

$$\mathcal{L} = \mathcal{L}_S + \lambda \mathcal{L}_C, \tag{1}$$

where $\mathcal{L}_S$ and $\mathcal{L}_C$ respectively denote classification loss and center loss, and λ is a trade-off parameter to balance the contribution of each component. We describe the motivation for the two loss terms in the following.

B. Classification Loss

Conventional supervised classification training requires *image-label* pairs, while labels are unavailable for generated data from a GAN model. In order to use the synthetic data for training, we propose two schemes to provide pseudo-labels for unlabeled data. Our empirical results show that both approaches improve the person re-identification performance. We first provide notations and brief background and then elaborate on the approaches.

For a single input image, the convolutional neural network calculates its feature representation $\mathbf{x}$ and output for k-th pre-defined class y_k, where $\mathbf{y} = \mathbf{W}^T\mathbf{x} + b$. Its estimated softmax probability to be classified into class k can thus be given by:

$$p(y_k) = \frac{e^{y_k - y_{max}}}{\sum_{j=1}^{K} e^{y_j - y_{max}}}, \quad s.t.,\ k \in [1, K], \tag{2}$$

where y_{max} represents the maximum response in $\mathbf{y}$, K is the number of pre-defined classes i.e., pedestrian identities in re-ID task.

1) One-hot label: One simple strategy is to assign a one-hot pseudo-label same as the real label following Fig. 1(b). Referring to the clustering criterion and considering the similarity between GAN generated and real image representations, we propose a straightforward yet effective solution to associate an unlabeled data sample to the most similar class.

The similarity between an input data representation in feature space $\mathbf{x}$ and center $\mathbf{c}_k$ for class k is formulated as below:

$$sim(\mathbf{x}, \mathbf{c}_k) = \frac{\mathbf{x} \cdot \mathbf{c}_k}{\|\mathbf{x}\|\|\mathbf{c}_k\|} \tag{3}$$

where $\mathbf{c}_k$ is the class specific anchor in the feature space and updated on the fly, whose detailed introduction can be found in Sec. IV-C. Pseudo-label ℓ for $\mathbf{x}$ is defined using the above mentioned similarity metric as follows:

$$\ell = \arg\max_k \ sim(\mathbf{x}, \mathbf{c}_k) \tag{4}$$

One advantage of one-hot pseudo-labeling is that it is consistent with ground truth labels, which enables unlabeled data to be integrated with labeled data for training without a separate training procedure with different loss formulation. They can be trained by following categorical loss function:

$$\begin{aligned} \mathcal{L}_S &= -\log(p(y_\ell)) \\ &= -(y_\ell - y_{max}) + \log(\sum_{j=1}^{K} e^{y_j - y_{max}}) \end{aligned} \tag{5}$$

The backward gradients can be written as:

$$\mathcal{L}_S^{'} = \frac{e^{y_\ell - y_{max}}}{\sum_{j=1}^{K} e^{y_j - y_{max}}} - 1. \tag{6}$$

2) Distributed label: The synthetic images generated by the GAN are random samples drawn from the approximated data manifold. Due to the complexity of high dimensional visual data, the pedestrian samples generated by GAN can have vague or absurd appearances and body shapes. These badly generated images succeed to pass the discriminator test yet they are easily distinguishable from the true samples when inspected by a human. Hence, it is inappropriate for an optimal learning procedure to arbitrarily consider these images to belong to an existing identity and assign a one-hot label as discussed above. To this end, we propose to treat single unlabeled data approximately as a weighted combination of representations from different classes. This scheme is illustrated in Fig. 1(c)

Accordingly, the final distributed label $q(y)$ is defined by a softmax function on the similarities between $\mathbf{x}$ and all cluster centers c, formulated as follows:

$$q(y_k) = \frac{e^{sim(\mathbf{x}, \mathbf{c}_k)}}{\sum_{j=1}^{K} e^{sim(\mathbf{x}, \mathbf{c}_j)}}, \tag{7}$$

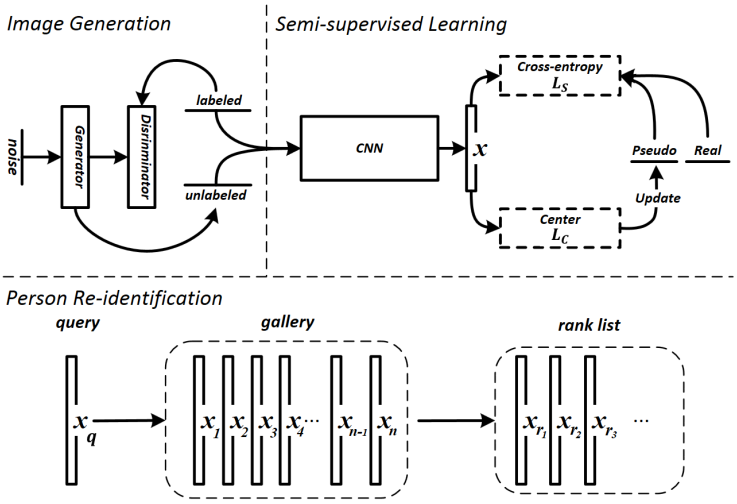

Fig. 2: The overall workflow of semi-supervised person re-identification. The top half denotes the training procedure while the bottom denotes the testing phase. Left part of training denoted as image generation consists of a GAN network that has one generator network and one discriminator network. The former takes in as input some random noise and outputs a generated sample. The discriminator tries to distinguish generated from real samples. The right half demonstrates the semi-supervised training on a CNN architecture with our proposed multi-task loss (dashed boxes). Center loss performs clustering of unlabeled data in feature space and simultaneously outputs pseudo-labels for them. During testing, we extract CNN output x as the pedestrian descriptor for both query and gallery images. Similar images are retrieved and ranked according to their descriptor similarity.

Distributed pseudo-label can thus be interpreted as the probabilities of unlabeled data belonging to each class. As suggested in [17], cross-entropy function can be used to train with distributed pseudo-label, for a single input, its classification loss is calculated as:

$$\begin{aligned}\mathcal{L}_S &= -\sum_{k=1}^{K} q(y_k)\log(p(y_k)) \\ &= -\sum_{k=1}^{K} q(y_k)(y_k - y_{max}) + q(y_k)\log(\sum_{j=1}^{K} e^{y_j - y_{max}})\end{aligned} \tag{8}$$

The corresponding gradients are written as:

$$\mathcal{L}_S^{'} = q(y_k)\frac{e^{y_k - y_{max}}}{\sum_{j=1}^{K} e^{y_j - y_{max}}} - q(y_k) \tag{9}$$

Specifically, if we constrain $q(y_k)$ to satisfy:

$$q(y_k) = \begin{cases} 1 & k = \ell, \\ 0 & k \neq \ell. \end{cases}$$

and plug it in Eq. 8 and Eq. 9, we can obtain Eq. 5 and Eq. 6, respectively. This means distributed pseudo-labeling scheme is a generalization of one-hot pseudo-labels, and both encodings derived from feature affinity can be used for training in a unified architecture with the original cross-entropy loss function.

C. Center Regularization

In addition to previous classification loss, we impose an extra center regularization term from [36]. This term associates each unlabeled sample to its matching cluster center in the feature space. It discovers underlying patterns in the feature space via center based clustering and thus performs intelligent data augmentation. We formulate the center regularization loss for a batch of m image feature representations $\{\mathbf{x}_i \in \mathbb{R}^d\}_1^m$ as follows:

$$\mathcal{L}_C = \frac{1}{2}\sum_{i=1}^{m} \|\mathbf{x}_i - \mathbf{c}_{\ell_i}\|_2^2 \tag{10}$$

where $\mathbf{c}_{l_i} \in \mathbb{R}^d$ is the center with dimension d of the cluster that $\mathbf{x}_i$ belongs to. For the case where an unlabeled data with distributed pseudo-label is being processed, $\mathbf{c}_{l_i}$ is selected using Eq. 4 for simplicity. This center regularization term is applied to **(a)** prevent large intra-class variations which can not be addressed by classification loss alone, **(b)** obtain centers of all categories which in turn help decide the label for unlabeled data, whether one-hot or distributed.

When losses are obtained, the backward gradients with respect to $\mathbf{x}_i$ can be calculated by:

$$\mathcal{L}_C' = \mathbf{x}_i - \mathbf{c}_{\ell_i} \tag{11}$$

Next, the cluster centers are updated using the following equation:

$$\Delta\mathbf{c}_k = \frac{\sum_{i=1}^{m} \delta(\mathbf{x}_i \in \mathbf{x}^L)\cdot\delta(\ell_i = k)\cdot(\mathbf{c}_k - \mathbf{x}_i)}{1 + \sum_{i=1}^{m} \delta(\mathbf{x}_i \in \mathbf{x}^L)\cdot\delta(\ell_i = k)} \tag{12}$$

where $\mathbf{x}^L$ is the set of labeled training data, and δ denotes delta function i.e., $\delta(condition) = 1$ if $condition$ is satisfied, and otherwise 0.

We train network with three notable modifications: (a) Instead of taking the entire training set into account, centers are updated based on mini-batches. (b) Only labeled samples in mini-batches are allowed to update class centers which proves to help stabilize the center constitution procedure since the pseudo-label may change dramatically for a same input between different training epochs. (c) The regularization loss is only back-propagated to labeled samples within each mini-batch.

The overall loss from Eq. 1 for a batch of input samples for both one-hot and distributed pseudo-labeling approaches can be rewritten uniformly as follows:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_S + \lambda\mathcal{L}_C \\ &= -\sum_{i=1}^{m}\sum_{k=1}^{K} q(y_k^i)\log(p(y_k^i)) + \frac{\lambda}{2}\sum_{i=1}^{m}\|\mathbf{x}_i - \mathbf{c}_{\ell_i}\|_2^2\end{aligned} \tag{13}$$

where y_k^i represents y_k for i-th input image.

The complete semi-supervised feature learning procedure is presented as Algorithm 1.

Algorithm 1: The semi-supervised feature learning with proposed pseudo-labeling approach

Input: labeled data: L, unlabeled generated data: U, maximum iteration: T, batch size: m, center update rate: α, trader-off parameter: λ, network parameters: θ

Output: Optimized parameters $\hat{\theta}$

Initialization: Training set $X = L \cup U$, Initialize θ with pre-trained ResNet-50, $\alpha = 0.5$, $\lambda = 10^{-4}$, class centers $\{\mathbf{c}_k = \mathbf{0} | k = 1, 2, ..., K\}$

for $t = 1 : T$ **do**
- Shuffle X and sample m samples to form a mini batch X^t;
- Feed forward X^t through CNN to obtain their feature representations $\mathbf{x}^t$;
- **for** $\mathbf{x}_i^t \in U$ **do**
  - Calculate $sim(\mathbf{x}_i^t, \mathbf{c}_k)$ using Eq. (3);
  - Generate pseudo label $\ell_{\mathbf{x}_i^t}$ for $\mathbf{x}_i^t$, one-hot with Eq. 4 or distributed with Eq.7.
- Compute the joint loss $\mathcal{L}^t$ using $\mathbf{x}^t$ with Eq. (13);
- **if** $\mathbf{x}_i^t \in L$ **then**
  - Update class center $\mathbf{c}^t$ with Eq. 12: $\mathbf{c}^t = \mathbf{c}^{t-1} + \alpha \Delta \mathbf{c}^t$.
- Backward propagation;
- Update the parameter set θ^t;

return: $\hat{\theta} = \theta^T$

D. Discussion

In this section, we study the interesting properties of our proposed method and compare it with existing works.

1) Feature Similarity vs Class Predictions: One significant difference between ours and all previous works on pseudo-labels generation is that this work is the first to propose assignment of pseudo-labels based on feature representation similarity in feature space.

One common practice of deep re-id works is that they first train an identity classification network and then extract last fully connected layer activations as the final descriptor to perform similarity calculation during the subsequent testing phase. Previous works such as [15, 18] calculate pseudo-labels based on classification prediction probability. The network predictions can be directly used for one-hot pseudo-labeling [18] if the class with maximum probability response is used as a label for training. However, probability fails in the case of distributed labels since pseudo-labels would be identical to class probability predictions which will resultantly not produce any weight corrections based on back-propagated gradients. Therefore, [15] proposed to rank the predicted probabilities and assign labels based on the ranking, which inevitably introduces inaccuracies. On the contrary, we propose to regard pseudo-label generation itself as a retrieval process with unlabeled data as query and labeled data as gallery based on representation similarity, which is identical to the final retrieval performed in person re-identification. In this way, the similarity based labeling scheme can derive pseudo-labels in both one-hot and distributed cases.

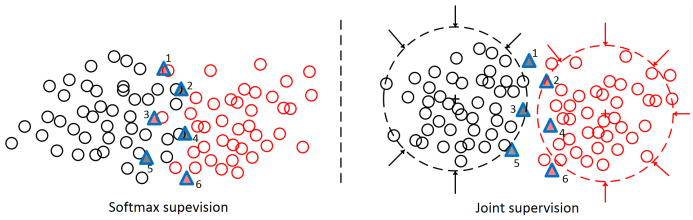

Fig. 3: This figure illustrates the distributions of different samples in the feature space before and after the center loss is imposed. Circles with different colors (e.g. red and black) stand for feature representations of *real (labeled)* samples form different categories. Blue triangles represent GAN generated *fake (unlabeled)* data. Generated data samples are denoted 1-6 from top to bottom. The filled color of each triangle denotes class label predicted under each case. Cross in each dashed circle denotes the center of that class. (Best viewed in color.)

2) Why centers matter?: The contributions of center regularization term are two-fold: *(a) It promotes learning more discriminative feature representations for labeled data.* Conventional classification loss only considers classifying samples correctly, the resulting deeply learned features therefore contain large intra-class variations. [36] proposed a center loss jointly with the softmax loss to improve the discriminative power of the deeply learned features by reducing intra-class variations. *(b) It produces pseudo-labels for unlabeled data considering their relationships with the labeled data.* This is an intuitive consideration because one can assume that the generated samples of a class are close to original ones. In contrast, previous works such as [18] and [15] inappropriately label the data with predicted probabilities and do not take into account their inherent relationships with the labeled data.

A toy example can be introduced to help illustrate above points. Left part in Fig. 3 shows a classification performed on a two-class (black and red) data set only with softmax loss. Unlabeled data samples (shown as blue triangles) are scattered around the boundary between two classes. Probability based labeling methods like [18] can be roughly seen as a nearest neighbor search. Sample 1 and 3 have as closest neighbor red points. Therefore, they are labeled red despite the fact in feature space they are more inclined towards black. Similar argument stands for data samples 2 and 4 being wrongly classified black. The right part is the labeling result after center loss is imposed. It is noticeable that less intra-class variations are introduced and unlabeled samples are more adequately classified considering the feature representation centers rather than the nearest neighbors.

3) Comparison with close work: The overall comparison of our approach with the closely related methods is summarized in Table I. We denote our two schemes as FAPL-o (one-hot) and FAPL-d (distributed), respectively. Existing strategies for labeling GAN data in person re-identification include all-in-one [13, 14], one-hot [18], LSRO [16] and dMpRL [15]. Their label distributions can be illustrated by Fig. 1(a), Fig. 1(b) and Fig. 1(c), respectively. Both LSRO and dMpRL adopts distributed labels, the difference is that LSRO selects uniform

TABLE I: Comparison with closely related work from different aspects.

Method	Label Assignment	Label Distribution	Label Source	Contributions on Pre-defined Classes
All-in-one[13]	static	one-hot	Manual	-
One-hot[18]	dynamic	one-hot	probability	-
LSRO[16]	static	distributed	Manual	same
MpRL[15]	dynamic	distributed	probability	different
FAPL-o	dynamic	one-hot	similarity	-
FAPL-d	dynamic	distributed	similarity	different

distribution while dMpRL considers ranking contributions.

Compared with [13, 16] which directly assign fixed and identical labels for all generated data, our proposed model considers the variations in-between them and dynamically predicts pseudo-labels in each iteration as the training progresses. [15, 18] assign labels dynamically, both of which adopt class probability predictions rather than feature similarity to assign labels. We propose to take advantage of feature similarities in the feature space and predict labels accordingly. Our experiments show that compared to the rigid *one-hot* class labels, *distributed* probability labels are more flexible and resilient.

In summary, our proposed approach enjoys the benefits of being more flexible, discriminative and aware of wide context in the feature space. As a result, it leads to better performance as evidenced through the reported quantitative comparisons in Section V.

V. EXPERIMENTS

In this section, we perform experiments on three widely adopted person re-identification datasets to evaluate the effectiveness of our proposed approach.

A. Datasets and Evaluation Protocol.

Market-1501 is a large-scale person re-identification dataset collected from 8 cameras on Tsinghua campus. In total it contains 12,936 images for training and 19,732 for testing, and the number of person identities for training and testing are 751 and 750, respectively. Overall, each identity in training set has 17.2 images on average. All pedestrian images are automatically detected by Deformable Parts Model (DPM) [37].

DukeMTMC-reID derives from a large multi-target, multi-camera pedestrian tracking dataset [38] and released by [16]. Pedestrian images in this dataset are captured by 8 cameras with hand-labeled bounding boxes. It comprises of 1,404 identities in which 702 are selected as training set and the rest 702 are used for testing. The training set contains 16,522 images which leads to an average of 23.5 images per training identity. The query set has 2,228 images of 702 identities from one camera to retrieve from the gallery with 17,661 images.

CUHK03 dataset is captured on the CUHK campus which contains 14,097 images of 1,467 identities. Two subsets are provided. One is hand-drawn annotated images and the other is detected by DPM [37]. We chose to use the detected set in our experiment. 9.6 images are obtained per identity in the training set. Since no standard training/test split protocol are provided, averaged results for 10 training/testing are reported.

Evaluation Metrics. We evaluate our method with rank-1 accuracy and mean average precision (mAP) on all three datasets. The rank-i accuracy denotes the rate at which one or more correctly matched images appear in top-i ranked images. The mAP value reflects the overall precision and recall rates, thus providing a more comprehensive evaluation metric.

B. Implementation Details

Re-id Baseline. In our experiments, we adopt the standard ResNet-50 proposed in [39] as the backbone architecture for our proposed approach. This network architecture has been used to evaluate closely related pseudo-labeling approaches, such as all-in-one[14], one-hot[18], and LSRO[16]. No other changes were made to the architecture for training expect for substituting the last 1000 class activation neurons to target identity number, i.e., 751 and 702 for Market-1501 and DukeMTMC-reID, respectively. We first resize all training images to be 256×256 followed by a random horizontal flipping and cropping to the input size 224×224. A dropout layer with 0.75 drop rate is inserted just before the final convolutional layer to prevent over-fitting for all datasets. The whole model is optimized by Stochastic Gradient Descent (SGD) with 0.9 momentum. The learning rate is set to 0.001 for 40 epochs and decayed to 0.0001 for another 10 epochs. During testing, last FC-layer with 2048-dim activations are extracted as the pedestrian descriptor for a cosine similarity based ranking. The network is implemented with the Matconvnet [40] package.

GAN Models. For fair comparison, we follow the training procedure in [16] and adopt DCGAN [21] as our model to generate fake unlabeled pedestrian images aiming at enlarging the training set. A 100-dim random is provided as the input to the **generator**, which is enlarged to form a $4\times4\times16$ tensor by a linear function, followed by the application of 6 deconvolutional layers in total with 5×5 sized kernels to obtain the desired $128\times128\times3$ image. The **discriminator** has 5 convolutional layers with 5×5 kernels to perform a binary classification task to separate real and fake samples from the given images. After the network is trained, we use the generator to produce up to 36,000 synthetic images. All synthetic data samples are resized to 256×256 for the following semi-supervised learning. Some generated data samples are displayed in Fig. 4, although some of the generated images are far from the actual data distribution, they still help regularize the model and improve performance. We present our experimental results in Section V-C.

Feature Extraction. We use ResNet-50 as the backbone network with our proposed loss formulation for feature learning. As a result, for an input image, the last layer activations from ResNet-50 are extracted as the pedestrian descriptor for retrieval. When tested on a NVIDIA TITAN Xp card, a single forward pass approximately takes 40 ms. Notably, when we use a batch of 100 inputs, the average extraction time for a single image is further reduced to 3ms.

C. Evaluation

1) The effectiveness of proposed approach: Our overall results are summarized in Table II. As shown, with the ResNet-50 as backbone architecture, the baseline achieved 72.74%,

TABLE II: Comparison with current state-of-the-art pseudo-labeling methods for person re-identification, namely, LSRO and MdRL on Market-1501 and DukeMTMC-reID datasets. 24,000 generated data are incorporated for training.

Methods	Market-1501		DukeMTMC-reID	
	rank-1	mAP	rank-1	mAP
Baseline	72.74	50.99	65.22	44.99
LSRO[16]	78.21	56.33	67.68	47.13
dMpRL-II[15]	80.37	58.59	68.24	48.58
FAPL-o (Ours)	82.04	61.26	70.92	51.99
FAPL-d (Ours)	**83.43**	**63.23**	**71.90**	**52.25**

65.22% rank-1 accuracy and 50.99%, 44.99% in mAP on Market-1501 and DukeMTMC-reID, respectively. We observe that the performance on DukeMTMC-reID is relatively lower than that on Market-1501, which is due to the heavy occlusions in the Duke dataset which makes the identification task more challenging. The goal here is to study the performance trend on the two datasets in comparison to other state of the art pseudo-labeling approaches. In this experiment, we randomly selected 24,000 generated images for two datasets as auxiliary data for training, and observed that both schemes lead to significant performance improvements over baseline. Rank-1 accuracy on Market-1501 increased by a margin of 9.30% and 10.69% for two schemes, respectively. mAP also enjoyed an increase from 50.99% to 61.26% and 63.23%. Similar trend is observed on DukeMTMC-reID with an overall improvement of 6.19% in rank-1 accuracy and 7.13% in mAP.

2) Amount of unlabeled data: We use DCGAN to generate up to 36,000 images, from which we randomly pick subsets to evaluate how the total number of synthetic images incorporated for training influence the re-ID performance. Results are displayed in Table III. For our two labeling schemes, considerable performance increase in both rank-1 accuracy (9.38%, 10.69%) and mAP (11.32%, 12.24%) metrics was observed over baseline. However, an increase in the amount of unlabeled images above a threshold (i.e., from 12000 to 36000) failed to demonstrate considerable boost in performance and the final result fluctuates around 82% for one-hot and 83% for distributed, respectively. Similar trend is observed amongst all other pseudo-labeling methods. We speculate that this phenomenon is due to the inherent representation ability of generated images. These images are sampled from a specific learned distribution (manifold of real data), therefore simply increasing sample numbers does not provide any additional information to the model that benefits the final retrieval task.

3) Amount of labeled data: It is desirable to learn better representation with less amount of labeled data. To test our approach when labeled data is extremely limited, we perform experiments on a reduced training set where available labeled examples are roughly cut to half and one third of the total amount. We follow the following rules when composing these subsets: (a) Keep all samples for identities with less than 8 images; (b) Keep half or one-third samples form identities with more than 8 images and discard rest. We then obtain *half* subset with 7,106 training images and *one-third* subset with 4,200 training images. Results can be found in Table IV. One noticeable fact is that when less labeled data is used, more unlabeled data is needed to achieve best performance. Also, with labeled data reduced to a half and a third, performance dropped for the case of baseline model, from 72.74% to 66.98% and 57.45% in rank-1 accuracy, respectively. This result is expected since less supervision is provided when training set is reduced. However, with our approach, we can observe an improvement of around 10% in rank-1 and 11% in mAP over baseline for all rows. Specifically, with unlabeled data, the *half* model (rank-1= $\sim$77%, mAP=55%) managed to outperform the fully-supervised baseline (rank-1=72.74%, mAP=50.99%) by a significant margin of 5%. Also note that the bottom line of subset *third*, unlabeled (36,000) images are roughly 8.5 times larger than labeled (4,200) in amount, our model still performs best (rank-1=67.87%, mAP=43.54%), which is a promising result showing that our approach can be applied on much smaller datasets.

4) Parameter sensitivity: We also conduct experiments studying the sensitivity of the trade-off parameter λ in Eq. 13, whose results are presented in Table V. λ is by default set to 10^{-4} in our experiments. We tried other settings for this parameter (such as 10^{-3} and 10^{-5}) and noticed a decrease in the performance of both one-hot and distributed labeling. For one-hot scheme, rank-1 accuracy dropped from 82.04% to 80.82% and 81.18% and mAP from 61.26% to 60.24% and 59.42%, respectively. Similar trend was observed on distributed labeling. Our empirical analysis is that the choice of lambda used in our experiments roughly makes both loss terms comparable.

(a) Real

(b) DCGAN

(c) IWGAN

Fig. 4: Examples of real and fake generated images on Market-1501. (a) The top row shows the real pedestrian samples from training set. (b) The middle row shows generated images from DCGAN, the visual quality is relatively low from the perspectively of a human viewer, but they can help regularize the model. (c) The bottom row shows the generated images from IWGAN. Better visual qualities in human body shape and clothing can be observed.

TABLE III: Rank-1 accuracy (%) and mAP (%) on Market-1501 dataset with varying numbers of unlabeled training data. Best results amongst approaches are in bold whilst best results for each method with different number of unlabeled data samples are underlined.

# GAN images	All-in-one [13, 14]		One-hot [18]		LSRO [16]		sMpRL [15]		dMpRL-I [15]		dMpRL-II [15]		FAPL-o(Ours)		FAPL-d(Ours)	
	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP
0 (baseline)	72.74	50.99	72.74	50.99	72.74	50.99	72.74	50.99	72.74	50.99	72.74	50.99	72.74	50.99	72.74	50.99
12000	76.96	55.68	76.52	55.69	77.17	55.22	77.73	55.27	77.88	55.84	79.22	58.14	81.38	60.31	**83.28**	**61.68**
18000	77.40	55.59	77.95	55.04	76.96	55.28	77.73	55.05	78.36	56.21	79.81	58.31	82.10	62.31	**83.16**	**62.38**
24000	77.21	56.07	77.62	56.90	78.21	56.33	78.85	55.59	77.79	56.10	80.37	58.59	82.04	61.26	**83.43**	**63.23**
30000	77.17	56.19	77.95	56.54	77.46	55.40	77.82	55.76	78.65	57.15	79.16	57.69	82.10	61.42	**83.02**	**62.41**
36000	75.92	55.24	77.42	56.38	77.91	55.82	78.32	55.45	78.95	57.42	79.90	57.61	82.12	60.70	**82.30**	**61.92**
Perf. boost	4.66	5.20	5.21	5.91	5.47	5.34	6.11	4.77	6.21	6.43	7.63	7.60	9.38	11.32	**10.69**	**12.24**

TABLE IV: Results on Market1501 dataset with reduced labeled data subsets. This is trained with distributed pseudo-labels. Best performance for each reduction case is shown in bold.

number of images	All		half		third	
	rank-1	mAP	rank-1	mAP	rank-1	mAP
0(baseline)	72.74	50.99	66.98	43.71	57.45	33.22
12000	83.28	61.68	76.81	54.25	66.39	42.87
18000	82.16	61.68	77.02	54.48	65.23	41.22
24000	**83.43**	62.23	77.46	55.26	66.48	42.04
30000	83.02	**62.41**	**78.59**	**56.50**	66.69	43.50
36000	82.30	61.92	78.15	56.30	**67.87**	**43.54**

TABLE V: Sensitivity analysis for the trade-off parameter λ balancing the contributions of classification loss and center loss. Experiments are performed on Market-1501 with 24,000

lambda	one-hot		Distributed	
	rank-1	mAP	rank-1	mAP
0.001	80.82	60.42	81.05	61.18
0.0001	**82.04**	**61.26**	**83.43**	**63.23**
0.00001	81.18	59.42	82.31	62.31

5) Unlabeled data with different visual quality: To better discover the effects that generated images with different quality has on regularizing the model, we select a state-of-the-art GAN model to perform our evaluation. We choose the recently proposed Improved Wasserstein GAN (IWGAN) [23] for image generation. IWGAN has strong theoretical guarantees compared to DCGAN due to the use of Wasserstein distance measure as an adversarial training loss which provides faster and more stable convergence.

For the **generator**, we draw a 128-dim random noise vector and use five 3×3 residual (with skip connections) deconvolution layers to up-sample it to obtain $128 \times 128 \times 4$ feature maps followed by another 3×3 convolution layer to generate the final $128 \times 128 \times 3$ output sample. A **discriminator** takes as input $128 \times 128 \times 3$ images and passes it first through a convolution layer to obtain a $128 \times 128 \times 64$ intermediate presentation and then through another five residual down-sampling convolution layer to a 8192-dim representation followed by a binary classification similar to DCGAN to predict whether the input is real or fake. Similar to DCGAN image generation, output images are resized to 256×256 for our model training.

We compare generated samples form both GAN networks in Fig. 4. One can notice that the generated images from both DCGAN and IWGAN are not comparable with real images, but it is clear that the IWGAN images shown in the bottom row are visually better than DCGAN images shown in the middle row. DCGAN generated images have less diversity, contain considerable distortions as well as ambiguous limbs and body shapes. On the contrary, IWGAN better preserves human body shape and can generate more realistic samples with large color variations in clothing.

For each generation method, we randomly select varying numbers of fake images as an addition to our real training set and report results in Table VI. On Market-1501, under the same one-hot pseudo-labeling setting, DCGAN achieved 81.38% in rank-1 while IWGAN achieved 82.84%. In comparison, on DukeMTMC-reID, a maximum of 1.48% rank-1 accuracy increase is observed when 12,000 samples are used in the distributed setting. Overall, two conclusions can be drawn from this experiment: (a) When better visual quality synthetic images are used from an improved GAN model, the performance across both datasets is further boosted by a margin of 0.5%-1%. This improvement is relatively small because sufficient samples from both models are considered which reduces the impact of bad quality samples. (b) Distributed pseudo-labeling approach consistently outperforms the one-hot approach no matter which GAN model is adopted.

6) Comparison with other pseudo-labeling approaches: In this section, we compare the proposed approach with all four existing pseudo-labeling methods that we are aware of on Market-1501 dataset. The compared pseudo-labeling methods include all-in-one[13, 14], one-hot[18], LSRO[16] and MpRL[15].

Among all published works, LSRO[16] is the state-of-art proposing to assign uniformly distributed pseudo-labels for unlabeled data to regularize the model. While MpRL[27] is a very recent work by improving the distribution by considering class contributions and achieves very competitive results. Three different implementations, sMpRL, dMpRL-I and dMpRL-II, are provided in [15]. More specifically, the first sMpRL assigned fixed distributed labels throughout the whole training process, this is similar to LSRO except for class contributions are considered when labels are produced. Both dMpRL-I and dMpRL-II dynamically assign pseudo-labels to each generated samples, but differ in when unlabeled data are used for training. Generated data are used from the start point of training in dMpRL-I, while used after 20 epochs in dMpRL-

TABLE VI: Rank-1 accuracy (%) and mAP (%) results on Market-1501 and DukeMTMC-reID datasets when two GAN models, DCGAN and IWGAN, are adopted for unlabeled image generation. For each dataset, best performance for one-hot and distributed schemes are underlined and in bold, respectively.

# GAN images	Market-1501								DukeMTMC-reID							
	DCGAN [21]				IWGAN[23]				DCGAN[21]				IWGAN[23]			
	One-hot		Distributed		One-hot		Distributed		One-hot		Distributed		One-hot		Distributed	
	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP
12000	81.38	60.31	**83.28**	61.68	82.84	62.87	83.05	**62.60**	71.57	52.68	71.68	52.83	72.21	53.23	**73.16**	**54.96**
18000	82.10	62.31	82.16	61.18	82.17	62.50	**82.89**	**62.26**	70.38	51.87	71.32	52.88	71.98	53.34	**72.60**	**53.97**
24000	82.04	61.26	83.43	63.23	82.42	61.58	**83.84**	**63.41**	70.92	51.99	71.90	52.25	71.72	53.01	**72.35**	**54.00**
30000	82.10	61.42	83.02	62.41	82.66	61.62	**83.58**	**63.78**	70.47	52.54	72.38	53.71	72.48	53.35	**73.44**	**54.71**
36000	82.12	60.70	82.30	61.92	82.51	61.29	**82.78**	**63.43**	71.23	52.18	72.40	53.73	72.40	52.76	**72.85**	**53.85**

TABLE VII: Comparison with state-of-art methods on market-1501 dataset. Best and second best results are denoted as bold and underlined text, respectively.

Method	Market-1501	
	rank-1	mAP
Gate-reID (ECCV'16) [41]	65.88	39.55
SCSP (CVPR'16) [42]	51.90	26.35
DNS (CVPR'16) [43]	61.02	35.68
ResNet+OIM (CVPR'17) [44]	82.10	-
Latent Parts (CVPR'17) [45]	80.31	57.53
P2S (CVPR'17) [46]	70.72	44.27
Consistent-Aware (CVPR'17) [11]	80.90	55.60
Spindle (CVPR'17) [47]	76.90	-
SSM (CVPR'17) [48]	82.21	68.80
JLML (IJCAI'17) [49]	85.10	65.50
SVDNet (ICCV'17) [50]	82.30	62.10
Part Aligned (ICCV'17) [51]	81.00	63.40
PDC (ICCV'17) [52]	84.14	63.41
LSRO (ICCV'17) [16]	78.06	56.23
dMpRL-II (Arxiv'18) [15]	80.37	58.59
Baseline	72.74	50.99
Ours-o+DCGAN	82.10	62.31
Ours-d+DCGAN	83.43	63.23
Ours-o+IWGAN	82.66	61.62
Ours-d+IWGAN	83.58	63.78
Ours-d+IWGAN+re-rank	**86.07**	**77.64**

TABLE VIII: Comparison of state-of-art approaches on the DukeMTMC-reID dataset. Rank-1 accuracy (%) and mAP (%) are reported.

Method	DukeMTMC-reID	
	rank-1	mAP
BOW+kissme (ICCV'15)[53]	25.13	12.17
LOMO+XQDA (CVPR'15)[54]	30.75	17.04
LSRO (ICCV'17)[16]	67.68	47.13
dMpRL (Arxiv'18)[15]	68.24	48.58
Verif + Identif (TOMM'17)[34]	68.90	49.30
APR (Arxiv'17)[55]	70.69	51.88
ACRN (CVPRW'17)[56]	72.58	51.96
PAN (Arxiv'17)[57]	71.59	51.51
FMN (Arxiv'17)[33]	74.51	56.88
Bilinear Coding (Arxiv'18) [58]	76.20	56.90
SVDNet (ICCV'17)[50]	76.70	56.80
DPFL (ICCVW'17)[32]	**79.20**	60.60
Baseline	65.22	44.99
Ours-o+DCGAN	71.57	52.68
Ours-d+DCGAN	72.38	53.71
Ours-o+IWGAN	72.40	52.76
Ours-d+IWGAN	72.85	53.85
Ours-d+IWGAN+re-rank	79.04	**70.74**

TABLE IX: Comparison with state-of-art methods on CUHK03(detected) dataset. Rank-1 accuracy (%) and mAP (%) are reported.

Method	CUHK03	
	rank-1	mAP
Gate-reID (ECCV'16) [41]	68.10	58.84
LOMO+XQDA (CVPR'15) [54]	46.30	-
dMpRL-II (Arxiv'18) [15]	68.68	73.48
LSRO (ICCV'17) [16]	73.10	77.40
SVDNet (ICCV'17) [50]	**81.80**	84.80
Baseline	70.68	74.25
Ours-o+DCGAN	73.28	78.92
Ours-d+DCGAN	74.17	79.62
Ours-o+IWGAN	73.99	78.64
Ours-d+IWGAN	74.58	79.89
Ours-d+IWGAN+re-rank	80.73	**86.38**

II when the CNN network is relatively stable.

Table III also summarizes the results of the state-of-the-art pseudo-labeling on person re-identification. It is shown that LSRO[16] achieved best performance with rank-1=78.21% and mAP=56.33% on Market-1501 dataset with 24,000 generated data. An overall performance increase amongst three MpRL[15] implementations can be observed, with dMpRL-II achieving the best results rank-1=80.37% and mAP=58.59%. Our proposed one-hot labeling scheme outperforms best competitor dMpRL-II by a margin of 1.75% and 3.72% in rank-1 accuracy and mAP, respectively, and distributed scheme further improves the performance to 3.06% and 4.64%. This is reasonable since distributed labels consider similarity contributions from each class and are more suitable for GAN generated data. Another remarkable fact is that despite our proposed labeling strategies assign pseudo-labels to unlabeled data immediately as training starts, they both outperforms the dMpRL-II which only starts to produce labels with relative stable CNN network after several epochs. This comparison proves that our proposed methods is superior to all state-of-the-art pseudo-labeling methods.

7) Comparison with state-of-the-art methods: Our work is dedicated to better exploiting synthetic images to boost re-id performance rather than beating the state-of-the-art results, however, we still compare our proposed approach with state-of-the-art works on Market-1501, DukeMTMC-reID and CUHK03 datasets to show its competence. As shown in Table VII, our distributed labels with IWGAN images achieved rank-1=83.58%, mAP=63.78% on Market-1501 dataset, which is very competitive with many state-of-the-art methods except

for JLML (rank1=85.1%, mAP=65.50%) and PDC. The main reason why JLML outperforms by a margin of 2% is because JLML incorporates three extra networks focusing on different local areas compared to our single branch architecture. With a state-of-art re-ranking technique from [59], we observed a further boost of 2.5% in rank-1 and 13.86% mAP demonstrating that reciprocal relationships are encoded in our learned identity representations. On DukeMTMC-reID dataset (Table VIII), we achieved 79.04% rank-1 accuracy and 70.74% with re-ranking. DFPL slightly outperforms ours in rank-1 (around 0.2%) because it takes advantage of multiple networks with different input scales and imposes consensus learning to force representations from different scales to be close if they belong to the same identity, while ours only adopts a single network. However, our mAP achieved 70.74% which is 10% higher than DFPL (60.60%). On the CUHK03 dataset (Table IX), it is clearly seen that our proposed model outperforms two competitive data augmentation methods LSRO and dMpRL by a margin of 1.48% and 5.9% in rank-1, respectively. With a re-ranking technique added, our model achieves 80.37% in rank-1 and 86.38% in mAP, which is competitive compared to SVDNet (81.80% and 84.80%). Without re-ranking, performances amongst all three datasets are inferior to other state-of-the-art approaches, this is partly due to our choice of a simple backbone network and partly because our work is focused on addressing pseudo-labeling problem for training data augmentation rather than explicitly dealing with hard failure cases, such as occlusion, scales and misalignment.

D. Ablation Study

We provide ablative experimental results on Market-1501 to evaluate each component of our proposed approach. The network is under full supervision of labeled data for baseline and center-loss and turns into a semi-supervised case when unlabeled data is introduced.

TABLE X: Ablative experiments in terms of each component of our proposed approach on Market-1501. 24,000 unlabeled data are incorporated for semi-supervised learning.

Methods	rank-1	mAP	supervision
Baseline	72.74	50.99	full
Center	79.45	57.25	full
FAPL-o	82.04	61.26	semi
FAPL-d	**83.43**	**63.23**	semi

1) Center loss: Center loss plays a crucial role in our proposed pseudo-labeling approach by reducing intra-class variations between data points, thus leading to more discriminative feature representations [36]. In this experiment, only labeled data is used for training to show the effectiveness of center regularization. When the center loss was applied (second row in Table. X), the baseline experienced a 6.71% rank-1 increase from 72.74% to 79.45% and 6.26% gain in mAP from 50.99% to achieve 57.25%. This confirms the positive effect center loss has on learning more discriminative representations.

2) Pseudo-labeling: On top of center loss, we add in our pseudo-labeling with both proposed schemes, denoted as FAPL-o and FAPL-d, respectively, in Table X. It can be observed that when synthetic data is incorporated for training with pseudo-labels, the network gains a further performance boost on both metrics to achieve 82.04% rank-1 accuracy for one-hot and 83.43% for distributed, 61.26% in mAP for one-hot and 63.23% for distributed, respectively.

VI. Conclusion

In this paper, we emphasize on the fact that a reasonable labeling approach for GAN generated images should consider representation similarity and encode their relationships with real data samples. To this end, we proposed a Feature Affinity based Pseudo Labeling (FAPL) approach with one-hot and distributed label encodings for the person re-identification task. Unlabeled images are assigned label encodings according to their distance to identity centers in feature space and help address re-id problem in a semi-supervised manner. Experiment results show that our proposed approach outperforms other pseudo-labeling methods on person re-identification task by a large margin and achieves competitive accuracy compared to state-of-the-art solutions.

References

[1] X. Wang, T. Zhang, D. R. Tretter, and Q. Lin, "Personal clothing retrieval on photo collections by color and attributes," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 2035–2045, 2013.

[2] N. Liang, G. Wu, W. Kang, Z. Wang, and D. D. Feng, "Real-time long-term tracking with prediction-detection-correction," *IEEE Transactions on Multimedia*, vol. PP, no. 99, pp. 1–1, 2018.

[3] Z. Liu, Z. Lin, X. Wei, and S. C. Chan, "A new model-based method for multi-view human body tracking and its application to view transfer in image-based rendering," *IEEE Transactions on Multimedia*, vol. PP, no. 99, pp. 1–1, 2017.

[4] K. H. Lee and J. N. Hwang, "On-road pedestrian tracking across multiple driving recorders," *IEEE Transactions on Multimedia*, vol. 17, no. 9, pp. 1429–1438, 2015.

[5] F. Chen, C. D. Vleeschouwer, and A. Cavallaro, "Resource allocation for personalized video summarization," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 455–469, 2014.

[6] Z. Wang, R. Hu, C. Liang, Y. Yu, J. Jiang, M. Ye, J. Chen, and Q. Leng, "Zero-shot person re-identification via cross-view consistency," *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 260–272, 2016.

[7] N. O'Hare and A. F. Smeaton, "Context-aware person identification in personal photo collections," *IEEE Transactions on Multimedia*, vol. 11, no. 2, pp. 220–228, 2009.

[8] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv preprint arXiv:1610.02984*, 2016.

[9] Y. Huang, H. Sheng, Y. Zheng, and Z. Xiong, "Deepdiff: Learning deep difference features on human body parts for person re-identification," *Neurocomputing*, vol. 241, pp. 191–203, 2017.

[10] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, and X. Xue, "Multi-scale deep learning architectures for person re-identification," *arXiv preprint arXiv:1709.05165*, 2017.

[11] J. Lin, L. Ren, J. Lu, J. Feng, and J. Zhou, "Consistent-aware deep learning for person re-identification in a camera network," in *CVPR*, 2017.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[13] A. Odena, "Semi-supervised learning with generative adversarial networks," *arXiv preprint arXiv:1606.01583*, 2016.

[14] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *NIPS*, 2016.

[15] Y. Huang, J. Xu, Q. Wu, Z. Zheng, Z. Zhang, and J. Zhang, "Multi-pseudo regularized label for generated samples in person re-identification," *arXiv preprint arXiv:1801.06742*, 2018.

[16] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," *ICCV*, 2017.

[17] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016.

[18] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on Challenges in Representation Learning, ICML*, 2013.

[19] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[20] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *arXiv preprint arXiv:1703.10593*, 2017.

[21] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[22] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017.

[23] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *NIPS*, 2017.

[24] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling, "Semi-supervised learning with deep generative models," in *NIPS*, 2014.

[25] G. Doquire and M. Verleysen, "A graph laplacian based approach to semi-supervised feature selection for regression problems," *Neurocomputing*, vol. 121, no. 18, pp. 5–13, 2013.

[26] M. Zhang, J. Tang, X. Zhang, and X. Xue, "Addressing cold start in recommender systems: a semi-supervised co-training algorithm," in *International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2014.

[27] S. Zhu, X. Sun, and D. Jin, "Multi-view semi-supervised learning for image classification," *Neurocomputing*, vol. 208, pp. 136–142, 2016.

[28] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *CVPR*, 2016.

[29] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamese long short-term memory architecture for human re-identification," in *ECCV*, 2016.

[30] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *ICPR*, 2014.

[31] S. Zhou, J. Wang, R. Shi, Q. Hou, Y. Gong, and N. Zheng, "Large margin learning in set to set similarity comparison for person re-identification," *IEEE Transactions on Multimedia*, vol. PP, no. 99, pp. 1–1, 2017.

[32] Y. Chen, X. Zhu, and S. Gong, "Person re-identification by deep learning multi-scale representations," in *ICCVW*, 2017.

[33] G. Ding, S. Khan, Z. Tang, and F. Porikli, "Let features decide for themselves: Feature mask network for person re-identification," *arXiv preprint arXiv:1711.07155*, 2017.

[34] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned cnn embedding for person reidentification," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 1, p. 13, 2017.

[35] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," in *ECCV*, 2016.

[36] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *ECCV*, 2016.

[37] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[38] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *ECCVW*, 2016.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[40] A. Vedaldi and K. Lenc, "Matconvnet: Convolutional neural networks for matlab," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 689–692.

[41] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *ECCV*, 2016.

[42] D. Chen, Z. Yuan, B. Chen, and N. Zheng, "Similarity learning with spatial constraints for person re-identification," in *CVPR*, 2016.

[43] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *CVPR*, 2016.

[44] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *CVPR*, 2017.

[45] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts

for person re-identification," in *CVPR*, 2017.

[46] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng, "Point to set similarity based deep feature learning for person re-identification," in *CVPR*, 2017.

[47] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *CVPR*, 2017.

[48] S. Bai, X. Bai, and Q. Tian, "Scalable person re-identification on supervised smoothed manifold," in *CVPR*, 2017.

[49] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," *arXiv preprint arXiv:1705.04724*, 2017.

[50] Y. Sun, L. Zheng, W. Deng, and S. Wang, "Svdnet for pedestrian retrieval," *arXiv preprint*, 2017.

[51] L. Zhao, X. Li, J. Wang, and Y. Zhuang, "Deeply-learned part-aligned representations for person re-identification," *arXiv preprint arXiv:1707.07256*, 2017.

[52] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *ICCV*, 2017.

[53] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015.

[54] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *CVPR*, 2015.

[55] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang, "Improving person re-identification by attribute and identity learning," *arXiv preprint arXiv:1703.07220*, 2017.

[56] A. Schumann and R. Stiefelhagen, "Person re-identification by deep learning attribute-complementary information," in *CVPRW*, 2017.

[57] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *arXiv preprint arXiv:1707.00408*, 2017.

[58] Q. Zhou, H. Fan, H. Su, H. Yang, S. Zheng, and H. Ling, "Weighted bilinear coding over salient body parts for person re-identification," *arXiv preprint arXiv:1803.08580*, 2018.

[59] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *CVPR*, 2017.